NWX-US DEPT OF COMMERCE

April 30, 2021
12:00 pm CT

Coordinator: Welcome and thank you for standing by. Today's call is being recorded. If you have any objections, you may disconnect at this time. All participants are in listen-only mode until the question-and-answer session of today's conference. At that time, questions will only be taken over the online Q&A feature. I would now like to turn the conference over to your host, Meghan Maury, from the US Census Bureau. You may now begin.

Meghan Maury: Thank you so much, Operator. Good afternoon and welcome, everyone. I'm Meghan Maury, a Senior Advisor here at the US Census Bureau, and I'm excited to be joining you for today's webinar on understanding the April 2021 demonstration data.

Today's webinar is designed to help our audience understand the demonstration data that the Census Bureau released this week, and to make sure you know how to navigate some of the tools we've provided to empower data users to analyze that demonstration data.

I'm joined today by Michael Hawes. Michael Hawes is part of a team that worked on building the 2020 Disclosure Avoidance System. His role has

been to take in user feedback, and make sure it's well reflected in how the algorithm is built.

I'm also joined by Matt Spence, who hopefully we will have on video in a moment, and Matt is part of our population division. He approaches the Disclosure Avoidance System, both as an internal data user, looking at how changes to the algorithm perform for the purposes of his own data use, but also is part of the team that's helping to provide the tools that let internal and external data users more easily see those - the changes to the algorithm reflected in the ultimate data that's produced. And there's Matt now.

I'll start off by grounding us in a little bit of shared language to make sure we're all on the same page throughout this webinar. Then Michael will take the mic and explain some of the changes that have been made to the system that you'll see reflected in the sample data.

Matt will then jump in and show you the tools his team has built to analyze the system. And Michael will give us a bit more information about how the privacy accuracy balance was set for this release. Finally, we'll reiterate the information about how you can provide feedback to the team, and we'll round out any time we have left, with questions from our audience.

So, let me just pause for a moment to talk about who this webinar is really designed for. Unlike some of the other webinars on Differential Privacy that you may have attended in the past, we're actually not covering what you might think of as Differential Privacy 101 today.

Instead, we'll be providing information that's most appropriate for people that have already spent some time engaging with the concepts of Differential Privacy, or Disclosure Avoidance Systems more generally. If that's you, we're

so glad that you're here. If you're not sure if that's you or not, we encourage you to stick around. We'll do our best to make the information we provide today, as accessible as possible.

If you are sure that's not you, and you feel like you need a Differential Privacy 101 before you're ready to engage with the demonstration data, we encourage you to take a look at the information about Differential Privacy on Census.gov, and/or to join us next week for a Differential Privacy 101 webinar. Once you feel more comfortable with that information, we hope you'll take the time to listen to the recording of this session so you can engage more deeply with the demonstration data.

Now, before we really jump in, let me cover a couple of technical pieces. First, I want to make sure to say that the views expressed in this presentation and by the panelists are those of the speakers and not the Census Bureau. Second, I really encourage you to type any questions you have in the Q&A box you'll see on your screen.

While we may not get to everyone's questions during today's session, we will take all of the questions that you submit today, and to either, again, answer them during this session, answer them during subsequent webinars, or follow up with you individually to make sure you have all the information you need. Finally, we'll be posting a video of this webinar. If you're watching the recording and have a question for the team, don't hesitate to email us at 2020DAS@census.gov.

With that, let's start with a tiny bit of terminology to make sure everyone understands the terms we're using in this webinar. First, I want to make sure that people understand that we're talking about Disclosure Avoidance Systems. The Disclosure Avoidance System is a tool we use to protect the

privacy of respondents by ensuring that the information they provide to the Census Bureau stays confidential.

It's kind of like an umbrella term that helps us describe what we're doing to protect confidentiality. Differential Privacy is actually a system of accounting that helps us measure the amount of privacy-loss associated with a particular disclosure avoidance tool.

I know that difference seems a little complicated. So, to put it another way, Differential Privacy helps us assess the risk of disclosure associated with the decennial census data after the TopDown Algorithm is applied, which I'm sure leads you to the question of, what on earth is the TopDown Algorithm?

The TopDown Algorithm is basically the algorithm or the formula we're using in 2020 to ensure that your data can't be re-identified. The TopDown Algorithm works a little bit like our other swap systems we've used in the past, like the swapping algorithm, but it's a much more complex tool than those more simplified systems were.

As our panelists speak today, you may hear a few other terms that are new to you, like PPMF or geographic spine. We'll try to explain those terms as we go, but if you hear something you don't understand, please feel free to use that Q&A feature, and we'll try to clarify for you today or in a follow-up conversation.

So now, let's get to the good stuff. Michael, we'll start with you. Your team has been engaged in this iterative process for a couple of years now to build this algorithm that will both protect the privacy of respondents and meet the accuracy needs of data users.

So, you build the algorithm, you put out some sample data, you get feedback, you make changes to the algorithm, put out more sample data, get more feedback and so on. Can you talk a little bit about the changes to the TopDown Algorithm that have come out of that process and how they're reflected in the data that our audience has been looking at this week?

Michael Hawes: Sure. Happy to. Thank you, Meghan. So, you're absolutely right that this has been an iterative process, and I'm sure it comes as no surprise to those attending today's webinar that census data are used for a wide variety of purposes, and our efforts to develop a system that can effectively protect the privacy of our respondents, while also ensuring high quality data to support those diverse use cases, would not have been possible without ongoing feedback and engagement from our data users.

Over the last couple of years, we've worked closely with members of our advisory committees, the Committee On National Statistics, American Indian Alaska Native Tribal leaders, professional associations, redistrictors, civil rights groups, researchers, state and local officials, and the list goes on and on.

And working with them, we've been able to get insights into what our data users need from the data in order for it to meet their use cases. And so, to enable all of those stakeholders to provide meaningful input into the development and improvement of our Disclosure Avoidance System, we've released five sets of demonstration data over the last year and a half.

And those demonstration data were generated by running 2010 census data through our TopDown Algorithm, so that data users can compare the new methods against the data that we originally published after the 2010 census. And with each set of demonstration data that we've released, the invaluable

feedback that we've received from the data user community and from their analyses, has really helped us to improve the algorithm.

And so, the latest set of demonstration data, which we released earlier this week, reflect the number of important changes that we made to the system based on that feedback from our data users. And Meghan, I think you - or Shelly, I think you have a graphic for me here. There we go.

So, as I mentioned, we've been getting feedback through a wide variety of forums and from a diverse range of data users and other stakeholders. Three of the pieces of information that we received a lot of feedback on led to direct important changes to the TopDown algorithm implementation, that are reflected in the demonstration data that we just released.

The first of these had to do with what we call off-spine geographies. So the TopDown Algorithm, and as its name suggests, starts at the nation level, and works its way down to smaller geographies. And it does this to ensure that as you examine data for larger and larger geographic entities, that the accuracy of the data will get better as you look at larger geographies, so that your state data will be incredibly accurate, and your county data will be accurate, your tract level data, slightly less accurate, and your block group level, slightly less, and so on, so that as you're looking at larger geographies, the accuracy increases. And that's a standard feature of official statistics all around the world.

The standard geographies that the TopDown Algorithm uses follow our normal tabulation geography. So nation to state, state to county, county to tract, tract to block group, and block group to block. But those are typically, or in many cases, not the geographies that our data users are particularly interested in.

They're interested in school districts, and places, and cities, and voting districts, and congressional districts. These assembled and constructed geographies that don't fall neatly on that kind of tabulation geographic hierarchy.

So, one of the things that we have been hearing from our data users, is that they wanted to see greater accuracy for these, what we call off-spine geographies, the ones that don't follow those normal tabulation geography boundaries.

So, one of the things that we changed in the recent set of demonstration data is, we implemented what's called an optimized geographic hierarchy for post-processing, or an optimized spine. And what this does is it essentially changes how the algorithm is processing down from the nation down to the block level, so that it's getting closer to those off-spine geographies that people are really interested in.

So, we essentially brought them closer to the spine, and we have more information about that in the documentation that came out with the demonstration data. And we'll be doing a separate webinar that goes into this much more deeply. But what this did was, it significantly improved the accuracy of the data for these non-tabulation geography entities like cities and places and school districts and so on. So, that was one change.

A second piece of feedback that we've heard a lot about from our data users is concern about outliers. So, if you've seen any of our prior information on our implementation of the Disclosure Avoidance System, you know that this is, at its core, a system that infuses noise into the statistics to protect privacy. It's introducing a little bit of uncertainty to prevent re-identification.

And that noise that's injected comes from a statistical distribution that's centered around zero. So, the most likely circumstancing for any statistic is that zero noise is added, but with slightly lower probability. You might have plus or minus one, with even lower probability plus or minus two, and so on. And it follows a statistical distribution.

In earlier iterations of the demonstration data that we released, we used a particular statistical distribution known as the Discrete Laplace Distribution, or Discrete Geometric Distribution. In response to concerns about outliers, where the larger amounts of noise were being added with low probability, but from the tail ends of that distribution, concerns about those outliers led us to change the mechanism that we're using for the noise infusion.

So, instead of using the Discrete Geometric Distribution, as we did in the past, we have transitioned to using a Discrete Gaussian Distribution. Now, this is a more complex statistical change here, but what it really means for the data is, it flattens out those tail ends of the distribution. So, it makes it significantly less likely that you'll get kind of outlier values of noise injected. And so, this change to the algorithm, does significantly decrease the likelihood of outliers in the resulting data.

And the third major change that we implemented for these demonstration data, addresses just overall concern by our data users about the level of accuracy. So, the first four sets of demonstration data that we released are all used to the same relatively conservative privacy-loss budget that was tuned more to the privacy side of the spectrum.

And we kept it there to enable us to identify kind of algorithmic problems that might be occurring, that might be distorting the resulting data. And so, we

held it there so that people could compare demonstration product to product over each release, to see the improvements that we were making to the algorithms.

In our most recent set of demonstration data that we released, we upped the privacy-loss budget. We essentially tuned for fitness for use, rather than tuning for privacy, as we had in the past. And what this has done is, it has significantly increased the overall accuracy of most of the statistics that are included in the demonstration data. And this new level of privacy-loss budget is much closer. It better approximates the overall level that we anticipate will be set for the actual production run of the 2020 redistricting file later this summer.

So, those are the three big changes that are reflected in the new demonstration data implemented because of the feedback that we've received. And we're looking forward to getting feedback from this new set of demonstration data so that we can continue to refine and improve the algorithm before we go into production. And so, I'll turn it back to Meghan. Thank you.

Meghan Maury: Thank you so much. That's really helpful. I know there are a lot of sort of smaller tweaks that are represented in those larger changes. And I know you all have done a ton of work to try to be responsive to the feedback that you've gotten from our data users, and I really appreciate it.

Matt, you're a person who is both sort of a data user, and someone who works for the Census Bureau, and you have a deep investment in privacy protection, so that you'll be able to collect complete data and comply with the law. I know that you've been doing a lot of analysis of how the Disclosure Avoidance System is functioning. Can you talk a little bit about how you're

analyzing the data and the tools you're providing to the public to help them do their own analysis?

Matt Spence: Absolutely. Thank you, Meghan. So, as you mentioned, I am in the Population Division of the US Census Bureau, and we are the sort of subject matter experts responsible for ensuring the quality of the data, the accuracy of the data throughout the entire process, from gathering the data, to publishing the data.

So, this is a natural extension of the work that we do already. We have since the beginning worked hand-in-hand with the disclosure avoidance science team and others to evaluate the accuracy of the data based on the use cases that we've received as part of these disclosure avoidance improvements and developments over the years.

So, I'm showing you now the webpage which talks about some of this demonstration data and the progress metrics that we published. We are going to go through in detail, the detailed summary metrics which our team of subject matter experts developed and added to in response to stakeholder feedback, which allows you to compare the accuracy, the bias, the fairness, and outliers, for these demonstration data.

So, if you scroll down, you can see that the latest release was the 28th of this month. And if you open that up, you'll see some information. We released two different versions of the data and the summary metrics with time. One uses the global epsilon of 12.2, which as Michael was saying, is much more likely to be closer to the level that is actually used for the redistricting data in August and September.

We also included a lower privacy-loss budget of 4.5, to allow users to

compare to prior runs of the demonstration data. So, you can see the improvements that are due to the algorithm, and the improvements that are due to the increased epsilon.

So, you can actually even download the micro data if you're interested, but right now, I'm going to be focused on these detailed summary metrics. So here, I'm going to open up the 12.2. When you open up the 12.2, you can see a little bit of a reading file. This discusses the improvements that have been made.

We have a list of all of the tables that are included as part of the summary metrics. Now, there's a little bit of a caveat here, in that some of these tables will be blank because we just simply don't have the data yet to publish data about this.

So one example is single year of age or sex. The redistricting data does not include single year of age or sex. They include voting age population, but they don't include the detailed age. We also don't include a lot of the housing unit data or family relationship data. Those data will be released in a later release as part of the Demographic and Housing Characteristics file (DHC), and those are coming later. But first, we're doing the PL.

So, the main tab of these demonstration metrics, these detailed summary metrics, is the third tab, what's called the updated BAP or Basic Accuracy Profile. And what you can see here is a list of tables with some metrics. These metrics generally fall into one of three categories.

First is measuring accuracy. That's very important. You know, obviously we want to see accurate data. So, how do we measure accuracy? Well, one way

that we do it is we compare - you know, we get this run from the Disclosure Avoidance System, the top-down algorithm.

We get a run of micro data, and we tabulate the data. And then we compare that - those tabulated data to those tabulated values to our actually published 2010 values. Remember, this is all 2010 data. So, this is really just looking to see how much the disclosure avoidance has changed, not actually seeing any changes over the last 10 years.

So, we're comparing 2010 data that's protected by the top-down algorithm to the published 2010 data. So, we measure that in a couple of ways. We can measure the accuracy in a couple of ways. So, the first one I'm going to start with is here, mean absolute error, MAE. And you can see here, this is the total population for counties.

We've got all counties here on those seven. Let me actually zoom in a little bit so everyone can see. I hope that's big enough. We got total population for counties. We've got an MAE of 4.91. Now, what does that mean in practice? Well, the way to think of it is that the average county will see a gain or a loss of about five people in their total population.

The precise way that this is calculated is, you take all 3,143 counties that were in the United States, excluding Puerto Rico, and you take the tabulated for each one of those. You take the tabulated data from the Disclosure Avoidance System, and you find the difference between that and the published data.

You take the absolute value of that. So, maybe you're off by 10, or maybe you're off by 20. And then you take the average across all 3,143, and that's the answer, 4.91. And that's the number of persons that we'll see on average in terms of change.

Another way to look at that is the MAPE, or Mean Absolute Percent Error. And this is a percentage of relative error. So, instead of looking at just the overall change, oh, you gained five people, you lost five people, you look at the relative change.

So, oh, this county gained half a percent. That county lost 1%. That county gained 2%, and you sum up all of that and pick the average. And so, what we see is that the MAPE is 0.04. So, that's a very small number. That's four one hundredth, of a percent, a very small number in terms of relative accuracy. So, very good. Zero is the goal for perfect accuracy.

There are some other columns as well, but one that I wanted to touch on also are counts of outliers here. We do have counts of outliers over here on the right. I'm going to come back to those, but I just wanted to show you sort of what we're looking at here.

The one feature of these detailed summary metrics, is that you can easily get a sense of how accurate the data were for this particular run. And you can also use it to compare across runs in the past. So, here on the screen, you should see the change in Mean Absolute Percent error. So, this is that relative error metric or MAPE for county total population.

And you can see that back in October 2019, our first demonstration data release, the result of error was about 0.7, 5%, and now it's 0.04%. So, what does that mean in actual persons? That went from 82 persons on average. The average county gained or lost 82 persons, and now the average county gains or loses about five people. So we're definitely seeing improvements there.

One of the other nice parts of these detailed summary metrics, is that we've broken it down by size category as well. We initially did this to see if there were sort of disproportionate impacts on the smaller geographies. And what we can see here is that the mean absolute error is approximately the same for all counties, despite the increase in size.

So, the smallest counties with a population of less than a thousand, gained or lost about four people, 3.75, and the average county of 100,000 or more people, gained or lost about five people. So, four to five people gaining or losing in absolute terms.

And so, that's a feature of the top-down algorithm, that geographies that are on spine will have approximately the same amount of error, no matter how big or how small the population is for that geography. So, counties are on spine, because nation, state, county. And because of that, the smallest counties and the largest counties will have approximately the same mean absolute error, despite whatever size that might be.

Now, let's look at incorporated places. Incorporated places are what many of us think of when we think of our hometown. These are cities and towns, political and legal entities across the US. In total, there were 19,540 in 2010. And we can do the same exact sort of calculation to find statistics about accuracy for incorporated places.

So here, we can see that the average incorporated place gained or lost about 21 people, which if you look at the MAPE, that's about 1.2% of the population. So, we can see that these places are slightly less accurate than the average county, but keep in mind that these places are frequently very, very

small. You can see that of those 19,000, 6,000 of them have fewer than 500 people.

So, those are going to have a larger percent error, but the larger incorporated places will have similar percent error to larger counties. One thing to note is, as long as I'm talking about sizes, is that unlike with counties, you'll see that places do - as they go up in population, they do tend to have a larger mean absolute error. Don't be alarmed. This is absolutely how it happens for these off-spine geographies.

Places frequently cross county lines, or cross tract lines. And so, they are not on the geographic spine. So, they will have this impact. But again, when you think about it in terms of the relative error, the MAPE, you can see that it's going down. It goes from 2.8 for those smallest places of fewer than 500 people, to 0.15, or the places with the largest population.

And once again, we can compare these across time. So you look at places now. This is place total population. We've got the mean absolute percent error here on the vertical access. You can see that in October of 2019, the average place saw a 10% change in its population, based on the top-down algorithm application. And now we're much lower. We're seeing it's below 2%. So it's gone, again, from about 85 persons the average place gained or lost, to now it's about 21 persons. So we are seeing improvements over time.

One thing I also wanted to touch on were the outliers. There was some discussion that the outliers were a concern, because it might be true that the overall number of counties or places might be improving, but maybe that meant that a couple of counties or a couple of places were getting much, much worse in terms of accuracy.

And so, we have these metrics that tabulate the number of the geographies that exceed a five percentage point error. How many of them gain more than 5% or lose more than 5% of their total population, or their population Hispanics or something like that.

And so here, this is, again, looking at incorporated places. We're looking at the number exceeding a five percentage point error for total population. And you can see that in October, it was nearly 8,000 of the 19,000.

So that's like 40% or so of the total number of incorporated places, had these very large errors of more than 5%. And now that number is about 700. So we've seen a significant increase. And a lot of that may be due to the improvements that Michael discussed in the algorithm and choosing the discrete Gaussian mechanism, which has those narrower tails, which means fewer very large outliers.

Those outlier metrics generally in these tables over here on the right. Here's that 700 number I was mentioning before, that there are 700 of those that have an absolute percent difference exceeding 5%.

Meghan Maury: This is such helpful information, Matt. I'm really - it's really giving me a sense of how to look at these charts and kind of draw my own conclusions about how the algorithm is performing on the sample - on the demonstration data at this time.

I wonder if you could dig with me even a tiny bit deeper and tell me - let's say I'm a local person. I'm trying to think about how to plan for the needs of my local Department Of Education, needs for funding, meeting the needs of those

marginalized folks in my community, planning for how many students there will be in our school in the next year, next five years down the road.

I know some of that I can't see yet until we have age data by year, which comes in later data products. I can only make an initial analysis, but I think there is some information I can see from these detailed summary metrics, that would help me make that initial analysis. Can you walk me through what I'd want to look at in here and how I might find it?

Matt Spence: Yes, absolutely. These detailed summary metrics are very, very extensive. You can see the list of tables really does extend to into hundreds of tables. The best thing we do is probably go on to the list of tables and search for a potential geography that you might be interested. Maybe it's school districts. Maybe it's Alaska native village statistical areas.

The other thing you can do is just sort of scroll through and see if any of these sort of call out to you. So, you mentioned school districts. School districts happen to be one of these ones that we've tabulated. We can scroll down here. Here, you can see we have separate tabulations for Puerto Rico. And here, here we are at school districts.

So now, we can evaluate elementary school districts, secondary school districts, and unified school districts, separately to see sort of what kind of error can we for these different levels.

So, just starting at the top, these elementary school districts, there are 2,300 of them across the US. Because they're off-spine, you'll see that the mean absolute error does increase as the size of the elementary school district increases. But in general, as before with places, the relative error goes down the larger you get.

So, the smallest school districts, elementary school districts, have a MAPE of 4.2%. So, their average total population changes by about 4.7%. And then the largest, with 100,000 or more in the elementary school district, they change by about 0.2%. And this is, again, in terms of total population.

And you can look and see secondary school districts as well, or unified school districts. You might be interested in minor civil divisions. There are several states that where minor civil divisions are a prominent political or legal feature. So those might be of interest.

We've also paid some attention to federal American Indian reservation and all preservation trust lands. We can see the mean absolute errors and MAPEs for those, for Oklahoma tribal statistical areas, for Alaska native village statistical areas, et cetera.

Now, one thing I'll note is that as you're scrolling through, you may come across a table that has a dash. I saw that up here somewhere. The dash indicates that it is zero.

But if it's blank, especially if the entire table is blank, let me scroll down and find one of those for us, yes, there. I believe there was a question actually about five-year age groupings. So, we can see here that the entire table is blank.

Well, as we mentioned, this is just the PL (P.L. 94-171 redistricting) data. This does not include that detailed age, and it doesn’t include sex. So, we don't have the data to be able to calculate the accuracy metrics, but once we do, and once we put out those demonstration data products for the DHC data, we'll be filling these in. And so, you'll be able to measure the accuracy there.

Right. So I do think that these detailed summary metrics are a great start. You can really dive into some of the very small characteristics data that we've got. We've really made an effort to incorporate stakeholder feedback to say, what is it that people would like to see as they're measuring the impact?

Here for example is one that is about voting age by Hispanic and by race alone. So here at the tract level. So we're really getting into the very small characteristics and very small geographies.

So we see that of the 73,000 tracts, the mean absolute error for the Hispanic White alone was four. And the mean absolute error for the non-Hispanic White alone was five. And you can compare that to prior runs and see just how much improvement there has been or you can say, well, how much needs to be done still.

Meghan Maury: And Matt, if I want to get even deeper, if I want to actually dig into the data itself or create other metrics that I think will help me analyze this data more effectively, is there a way to do that?

Matt Spence: Yes. Thanks to our friends at IPUMS and David Van Riper, they have tabulated the micro data for us. And so this is on the IPUMS website. It's actually the NHgis.org website. You can see that they have released all of these prior PPMFs. But in this one, in this case, this is the 2021-04-28 version.

So I'll scroll back down to those. Here we go. And so now, what you can do is you could say, okay, I'm interested in seeing the actual experience. It's one thing to see summary metrics, which talk about overall measures of accuracy, but I want to see how my particular place is affected.

Rather than go and download a 15 gigabyte text file or SAS file or do the tabulation, you can just go ahead and use the NHgis.org IPUMS website to pull down, what is my county's population? And you pull down county, it will actually - it's a really nice tabulated dataset that has total population, as well as all the various characteristics that you'd be interested in. And it's right there, right next to the published 2010 data from the SF 1.

So you can easily see, well, my county has 50,000 people, and what was it published? Well, it was published 49,000 people. So, I can see that I've gained 1,000 people. You can also look at legislative districts, congressional districts, core-based statistical areas. You can even go down to the block group level or block level, if you really, really want to get into the very detailed geographies.

Meghan Maury: Thank you. That's so exciting and interesting, especially for those of us that self-identify as nerds and really want to dig in there. Let me just change face for a moment. Matt, you've been talking a lot about error rates so as we walk through these metrics.

And I guess that brings up the question of how we thought about how to assess the level of error that's allowable in this system. And of course, on the flip side, how we're ensuring that enough privacy is inherent in the system to protect respondents' information.

Michael, I know you were a key part of the team that was thinking that through. Can you talk about how you came to a decision about what level to set that privacy accuracy dial at, and what it means for the main use cases of the dataset that will be released in August and September?

Matt Spence: Absolutely. And it's a difficult balance, because you need to make sure that you're introducing enough uncertainty to protect individuals' privacy and the confidentiality of their data, while also ensuring that the data are accurate enough for their intended uses.

So, I already talked about many of the changes that we made to the algorithm itself that led to accuracy improvements. But in terms of setting the privacy-loss budget, that point on the spectrum between perfect privacy on one side and perfect accuracy on the other, we needed some sort of reference point.

We needed kind of a concrete like fitness-for-use standard to evaluate against. And so, the way we developed that for the tuning runs that we were doing, was to start with the primary use cases for the PL 94-171 redistricting data, and that's for the purposes of redistricting, and for enforcement of the Voting Rights Act.

And so, with those statutory use cases at the forefront, we conferred with attorneys at the Department of Justice's Voting Rights Section, and with redistrictors, and civil rights groups, and others, on what their concerns were regarding accuracy for this bigger product.

And based on the feedback that we got from those data users, we established the standard that we would use for the experimental tunings that we were doing for this demonstration data release. And so, kind of the key things that we learned from that feedback was that for these statutory use cases, kind of the key data elements that people were interested in, were total population counts, total voting age population counts, and the proportions of racial groups for un-pre-specified geographies.

So, for moving districts, for congressional districts, for precincts, for wards,

for towns, that we may not know what the boundaries are until after the census, these off-spine geographies that I was referring to earlier. So, with those data elements and geographies in mind, we established tuning targets that we ran to ensure that we would be able to meet sufficient accuracy for both on spine and off-spine geographies for those data elements.

And one of the specific tuning targets that we used was to ensure that, for any on spine or off-spine geography that has at least 500 people in it, we wanted to ensure that the largest racial group for that geography has a proportion of the total population for that year.

I mean, that largest racial group will be within five percentage points of the enumerated value at least 95% of the time. And so, that was one of the driving accuracy targets that we used for that tuning that led to the production of this set of demonstration data.

Now, in reality, we actually far exceeded the targets that we set. As I said, we were looking for that kind of five percentage points in the proportion of the largest group, at least 95% of the time for on spine, and off-spine entities that have populations of between 500 and 549 people. So, very small voting district size entities.

In reality, for that kind of size level for the 500 and 549, we ended up hitting those targets 99.5% of the time. So 99.52% of the time for off-spine places and other off-spine entities. And then as those areas get larger, as the population size gets larger, those with larger populations actually perform even better against those standards.

So, those were the targets that we set. We're eager to have our data users go in and evaluate the demonstration data at that level of accuracy and get back

to us and, does it meet your use case fitness for use needs, or do we need to run a little additional tuning to better make it meet those needs?

Meghan Maury: Yes. Thank you for that. And just a follow-up question on that, Michael. I know Matt just walked us through all these other ways of looking at quality. And you talked about a bunch of other changes to the algorithm. Just for clarification, those accuracy targets, was that the only thing that you were looking at when you were looking at sort of how the algorithm was performing, or was it kind of a broader spectrum?

Michael Hawes: Oh, those were not the only things we were looking at by any stretch of the imagination. We were looking at a wide array of accuracy measures, many of which Matt walked through in his presentation before. Those were some of the primary ones that we used for the geographical tuning of the algorithm certainly, because those reflect the statutory use cases that we know we need to be considering.

But we were also looking at an assortment of racial statistics, Hispanicity statistics, looking at population counts at various levels of geography, all with an eye towards, like, are we getting sufficient accuracy for that level of privacy protection?

Also, recognizing that the data that are going to be included in the PL file, will also then be reflected in subsequent data products. So, we have to make sure that the level of accuracy that we're establishing now will support the level of accuracy that's going to be needed in those subsequent data products.

Meghan Maury: Got it. Thank you for that. And we're about to turn to answering a little bit of the Q&A, although we've been doing our best to answer them in real time in

the Q&A feature. Before we jump in there, I just wanted to draw folks' attention to a couple of other things.

You'll see a slide up right now for a series of webinars, much like this one, where we'll dig in deeper on different topics about the Disclosure Avoidance System. And I think some of the more technical questions I see in the Q&A feature, may be best addressed in those sessions, where we can really get into the weeds with you all.

I also wanted to make sure that folks knew - and again, you can access that - the links to all of those webinars very soon. We'll post links in the update section of our website and in the newsletters. We'll also be archiving recordings to those on the website.

So, don't feel like if you can't make this particular day or time, that the information is not going to be available to you. You can tune in later, and never hesitate to ask us questions. Again, I will repeat the email address where you can kind of reach us directly, which is 2020das@census.gov.

I also want to make sure that folks are aware of how to submit more formal feedback to us on the Disclosure Avoidance System itself. Here's some information about how to submit feedback. The changes in the PPMF were pretty vast. We know people are going to want to provide their own analysis to this - to the sample data.

And we'd love to hear anything that you want to tell us about how the algorithm is performing. The input you give us will help inform the final decision-making about where that privacy accuracy budget is set, how the parameters work within the system.

We'd love to get your feedback by May 28th, and you can send feedback to us through 2020das@census.gov. I know you're getting tired of me saying that. And we just put some prompts here of things that would be especially helpful for us.

Feedback on fitness for use, like whether the data is working for your particular use case, how you came to that conclusion. If it's not working well for you, are there changes, are there tweaks we can make to the algorithm that would help it better perform for your use case? Have you seen any of these improbable results or the outliers that would be helpful for us to understand?

We'd also love feedback on the privacy side of the question. Did the proposed products present any confidentiality concerns that you want to make sure that we're addressing in the deck? And finally, feedback about improvements.

Are there improvements that you've identified in how the algorithm is performing, that you want to make sure that we retain in that final decision-making? Particularly if there are types of geographies, error metrics where you're seeing significant improvement, you want to make sure that that doesn't get lost in the final decision-making, we want to hear about that too.

I know that you've got a couple of other slides that show people how to stay informed, where to get more information. We can throw those up as well. And then I'm seeing a bunch of questions in the chat.

I know many of them have already been addressed, but Michael, I'm seeing quite a few about the accuracy target. I don’t know if you can see them in your window, but people are asking about the quality metrics, sort of more definition around how the geographies function. Are there differences

between different race and ethnicity groups? Is there more information that you have about what happens outside that 99%?

Michael Hawes: So, that's a great question, and I don't have that data in front of me. We are actually conducting a much more extensive empirical analysis of the fitness for use of the current algorithm parameters for the redistricting and Voting Rights Act use cases.

We'll be releasing that empirical analysis in the next few weeks, and one of the webinars that Meghan mentioned will be specifically focused on that empirical analysis. So, I would say, it is something we are looking at. We are evaluating it, and stay tuned and we'll be providing that information.

Meghan Maury: Yes. Thank you for that. I do hope people will come back for the other webinars in this series. I also saw a couple of questions in the chat, which I think Michael, you answered, but I just want to make sure that we get this sort of out to everyone listening.

I know a lot of people are curious about how we're building the Disclosure Avoidance System for the data products that are released further down the road, like the DHC, and DDHC, the detailed demographic information that many people use for modeling and tons of other purposes.

Have there been any decisions made so far, Michael, that will impact how that process works for DHC and other later data products?

Michael Hawes: So we've been focusing a lot of our attention for the last seven months on getting the PL data tuned and getting ready for production of the PL data this summer. And that has required delaying some of the major decision-making for the demographic housing characteristics file.

But we are now turning our attention to that, and beginning the stakeholder engagement process that will allow us to do the same kind of tuning of the algorithm for the DHC use cases that we have all - that we've done for the redistricting use cases.

That process is ongoing. We will be releasing more information about timelines and duties or expectations, and we will be releasing demonstration data for the DHC. We will be doing extensive stakeholder engagement to make sure that we're developing and using appropriate accuracy targets for DHC use cases, and that will be occurring over the coming months.

Meghan Maury: Yes, thanks for that. And I think kind of implicit in what you're saying is about the decision-making for that process. We really want to be informed by the data users, by the stakeholders who are the ultimate users of these products. And so I think, from my understanding, the decisions haven't been made, because we want to get more feedback from the stakeholders before any of those final decisions are made. There are quite a lot more questions in the Q&A. I know we're just four minutes from the end.

Many of them are specific questions about specific data points in the detailed summary metrics. I think Matt is trying to answer those one by one, but Matt, is there anything else you want folks to know? I know you've taken a look at these Q&As, that would help them to know what they're looking at?

Matt Spence: I've seen a couple of questions about interpreting the size categories. And in general, people are getting it absolutely right. It's the top line of the table is all of the geographies of any size. And then the sort of sub rows, the rows underneath, split it up by size category.

So, it's counties under 1,000 people or, Federal American Indian Reservations over 10,000 people, or between 1,000 people. And so, those characteristics are specific to that set of geographies. That's the idea.

I know there were some other questions about how the largest demographic group was determined, the combination between Hispanic and race. And I was wondering if Michael could address those. Those are some that I cannot quite address.

Michael Hawes: So, unfortunately, I was in the process of reading questions when you spoke and I missed what you were asking me to do.

Matt Spence: Oh, no problem. I just see a couple of questions about - and maybe it's easy to discuss, a couple of questions about how Hispanic, sort of combined with race, for these target metrics for the redistricting use case.

Michael Hawes: Absolutely. So, for the purposes of the tuning, we were looking at a subset of race categories. So the Department of Justice, for the purposes of redistricting and Voting Rights Act enforcement, uses the OMB 11 category, race and Hispanicity variable categorization, which is - it's a mapping of the larger race and Hispanic origin categories that we actually collect on the census.

So, on the census itself, we collect 63 race categories, every possible combination of those race categories, crossed with Hispanic and non-Hispanic. So, that gives 126 possible outcomes there, which get mapped to the 12 that are used - the 12 OMB categories that are used by Department Of Justice.

So, the tuning was largely focusing on those 12 that factored Hispanic or non-

Hispanic into the race categories themselves. And that's what we were primarily focusing on with regards to that specific accuracy target. But we were absolutely looking at the broader 63 category race variable, and the Hispanic, non-Hispanic, crossed with that, as part of our broader evaluation of overall fitness for use of the data.

Meghan Maury: Yes, that's really helpful. I know there are a lot more questions in the Q&A. We will do our best to follow up with folks to get them answers to the questions that they've asked. Again, don't hesitate to email us with any particular questions, whether you've added them in the chat or not, at 2020das@census.gov.

I'll close us out with one final piece. Yvette, you asked about if we'll publish anything on how federal agencies might change how they use census data in funding formulas. To some extent, that is a question that will be answered a little bit further down the road.

Most of the data that agencies rely on for federal funding formulas actually come from those later data products like the DHC, DDHC, or even our population estimates, which are a little bit separate from this whole process.

So, we'll definitely be continuing to communicate with folks, especially through that stakeholder engagement process, about how the algorithm is functioning and what other opportunities for differences might exist outside this process as well.

I really appreciate everyone participating. I really appreciate your questions. I think we got a lot of good questions already that can help us inform how we develop the rest of the webinars in this series. So, I hope you continue to tune in.

The next webinar will be on May 4th, on Differential Privacy 101. More webinars will follow in the coming weeks. And again, if you have any questions at all, please don't hesitate to reach out to us by email. Matt or Michael, any last words before we sign off?

Michael Hawes: Just to thank you all for attending, and we're looking forward to being able to provide more information over the coming webinars.

Matt Spence: Thank you so much.

Meghan Maury: Thank you all so much.

Coordinator: Thank you for your participation in today's conference. You may disconnect at this time.

END